CHAPTER 5

Reliability of Physical Findings

KEY TEACHING POINTS

- *Reliability* refers to how often two clinicians examining the same patient agree about the presence or absence of a particular physical finding. Commonly used measurements of reliability are *simple agreement* or the *kappa (κ-) statistic*.
- About 60% of physical findings have κ-statistics of 0.4 or more, indicating that observed agreement is moderately good or better.
- Despite the common belief that technologic tests are more precise than bedside observation, the κ-statistics observed for most diagnostic standards (e.g., chest radiography, computed tomography, angiography, magnetic resonance imaging, endoscopy, and pathology) are similar to those observed for physical signs.
- Some causes of interobserver disagreement can be eliminated, but because clinical medicine is inherently a human enterprise (even when interpreting technologic tests), subjectivity and a certain level of clinical disagreement will always be present.

Reliability refers to how often multiple clinicians, examining the same patients, agree that a particular physical sign is present or absent. As characteristics of a physical sign, reliability and accuracy are distinct qualities, although significant interobserver disagreement tends to undermine the finding's accuracy and prevents clinicians from applying it confidently to their own practice. Disagreement about physical signs also contributes to the growing sense among clinicians, not necessarily justified, that physical examination is less scientific than more technologic tests, such as clinical imaging and laboratory testing, and that physical examination lacks their diagnostic authority.

The most straightforward way to express reliability, or interobserver agreement, is **simple agreement**, which is the proportion of total observations in which clinicians agree about the finding. For example, if two clinicians examining 100 patients with dyspnea agree that a third heart sound is present in 5 patients and absent in 75 patients, simple agreement would be 80% (i.e., [5 + 75]/100 = 0.80); in the remaining 20 patients, only one of the two clinicians heard a third heart sound. Simple agreement has advantages, including being easy to calculate and understand, but a significant disadvantage is that agreement may be quite high by chance alone. For example, if one of the clinicians in our hypothetical study heard a third heart sound in 10 of the 100 dyspneic patients and the other heard it in 20 of the patients (even though they agreed about the presence of the heart sound in only 5 patients), simple agreement by chance *alone* would be 74%.* With chance agreement this high, the observed 80% agreement no longer seems so impressive.

* Agreement by chance approaches 100% as the percentage of positive observations for both clinicians approaches 0% or 100% (i.e., both clinicians agree that a finding is very uncommon or very common). The Appendix at the end of this chapter shows how to calculate chance agreement.

To address this problem, most clinical studies now express interobserver agreement using the kappa (κ-) statistic, which usually has values between 0 and 1. (The Appendix at the end of this chapter shows how to calculate the κ-statistic.) A κ-value of 0 indicates that observed agreement is the same as that expected by chance, and a κ-value of 1 indicates perfect agreement. According to convention, a κ-value of 0 to 0.2 indicates *slight* agreement; 0.2 to 0.4 *fair* agreement; 0.4 to 0.6 *moderate* agreement; 0.6 to 0.8 *substantial* agreement; and 0.8 to 1.0 almost *perfect* agreement.[†] Rarely, physical signs have κ-values less than 0 (theoretically as low as –1), indicating the observed agreement was worse than chance agreement.

Table 5.1 presents the κ-statistic for most of the physical signs discussed in this book, demonstrating that, with rare exceptions, observed agreement is better than chance agreement (i.e., κ-statistic exceeds 0). About 60% of findings have a κ-statistic of 0.4 or more, indicating that observed agreement is moderate or better.

Clinical disagreement occurs for many reasons—some causes clinicians can control, but others are inextricably linked to the very nature of clinical medicine and human observation in general. The most prominent reasons include the following: First, the physical sign's definition can be vague or ambiguous. For example, experts recommend about a dozen different ways to perform auscultatory percussion of the liver, thus making the sign so nebulous that significant interobserver disagreement is guaranteed. Ambiguity also results if signs are defined with terms that are not easily measurable. For example, clinicians assessing whether a peripheral pulse is present or absent demonstrate moderate-to-almost perfect agreement (κ = 0.52 – 0.92, see Table 5.1), but when the same clinicians are asked to record whether the palpable pulse is normal or diminished, they have great difficulty agreeing about the sign (κ = 0.01 – 0.15) simply because they have no idea what the next clinician means by "diminished." Second, the clinician's technique may be flawed. For example, common mistakes are using the diaphragm instead of the bell of the stethoscope to detect the third heart sound, or stating a muscle stretch reflex is absent without first trying to elicit it using a reinforcing maneuver (e.g., Jendrassik maneuver). A third reason for clinical disagreement involves biologic variation of the physical sign. The pericardial friction rub, pulsus alternans, cannon A waves, Cheyne-Stokes respirations, and many other signs are notoriously evanescent, tending to come and go over time. Fourth, the clinician could be careless or inattentive. The bustle of an active practice may lead clinicians to listen to the lungs while conducting the patient interview or to search for a subtle murmur in a noisy emergency room. Reliable observations require undistracted attention and an alert mind. Lastly, the clinician's biases can influence the observation. When findings are equivocal, expectations influence perceptions. For example, in a patient who just started blood pressure medications, borderline hypertension may become normal blood pressure; in a patient with increasing bilateral edema, borderline distended neck veins may become clearly elevated venous pressure, or in a patient with new weakness, the equivocal Babinski sign may become clearly positive. Sometimes, biases actually create the finding: if the clinician holds a flashlight too long over an eye with suspected optic nerve disease, he may temporarily bleach the retina of that eye and produce a Marcus Gunn pupil, thus confirming the original suspicion.

The lack of perfect reliability with physical diagnosis is sometimes regarded as a significant weakness, leading to the charge that physical diagnosis is less reliable and scientific than clinical imaging and laboratory testing. Nonetheless,

[†] No measure of reliability is perfect, especially for findings whose prevalence clinicians agree approaches 0% or 100%. For these findings, simple agreement tends to overestimate reliability, and the κ-statistic tends to underestimate the reliability.

TABLE 5.1 Interobserver Agreement and Physical Signs

Finding (ref)	κ-Statistic*
GENERAL APPEARANCE	
Mental Status Examination	
Mini-mental status examination[1]	0.28-0.80
Clock-drawing test (Wolf-Klein method)[2]	0.73
Confusion assessment method for delirium[3-6]	0.70-0.91
Altered mental status[7]	0.71
Stance and Gait	
Abnormal gait[8,9]	0.11-0.71
Skin	
Patient appears anemic[10,11]	0.23-0.48
Nailbed pallor[12]	0.19-0.34
Conjunctival pallor (rim method)[13]	0.54-0.75
Ashen or pale skin[7]	0.34
Cyanosis[10,14]	0.36-0.70
Jaundice[15]	0.65
Loss of hair[16]	0.51
Vascular spiders[15-17]	0.64-0.92
Palmar erythema[15-17]	0.37-1.00
Hydration Status	
Patient appears dehydrated[10]	0.44-0.53
Axillary dryness[18]	0.50
Increased moisture on skin[10]	0.31-0.53
Capillary refill >3 s[7]	0.29
Capillary refill >5 s[19]	0.74-0.91
Nutritional Assessment	
Abnormal nutritional state[10]	0.27-0.36
Other	
Consciousness impaired[10]	0.65-0.88
Patient appears older than age[10]	0.38-0.42
Patient appears in pain[10]	0.43-0.75
Generally unwell in appearance[10]	0.52-0.64
VITAL SIGNS	
Tachycardia (heart rate >100/min)[20]	0.85
Bradycardia (heart rate <60/min)[20]	0.87
Systolic hypertension (SBP >160 mmHg)[20]	0.75
Hypotension (SBP <90 mmHg)[20,21]	0.27-0.90
Osler sign[22-24]	0.26-0.72
Rumpel-Leede (tourniquet) test[25,26]	0.76-0.88
Elevated body temperature, palpating the skin[10]	0.09-0.23
Tachypnea[7,14,20]	0.25-0.60

Continued

TABLE 5.1 Interobserver Agreement and Physical Signs—cont'd

Finding (ref)	κ-Statistic*
HEAD AND NECK	
Pupils	
Swinging flashlight test (relative afferent pupil defect)[27]	0.63
Diabetic Retinopathy	
Microaneurysms[28,29]	0.58-0.66
Intraretinal hemorrhages[28,29]	0.89
Hard exudates[28,29]	0.66-0.74
Cotton wool spots[28,29]	0.56-0.67
Intraretinal microvascular abnormalities (IRMA)[28,29]	0.46
Neovascularization near disc[28,29]	0.21-0.48
Macular edema[28,29]	0.21-0.67
Overall grade[28,29]	0.65
Hearing	
Whispered voice test[30]	0.16-1.0
Finger rub test[31]	0.83
Thyroid	
Thyroid gland diffuse, multinodular or solitary nodule[32]	0.25-0.70
Goiter[33,34]	0.38-0.77
Meninges	
Nuchal rigidity, present or absent[35-37]	0.24-0.76
LUNGS	
Inspection	
Clubbing[14,38] (general impression)	0.33-0.45
Clubbing (interphalangeal depth ratio)[39]	0.98
Clubbing (Schramroth sign)[39]	0.64
Breathing difficulties[10]	0.54-0.69
Gasping respirations[7]	0.63
Reduced chest movement[14,40,41]	0.14-0.38
Kussmaul respirations[42]	0.70
Pursed lip breathing[41]	0.45
Asymmetric chest expansion[43]	0.85
Scalene or sternocleidomastoid muscle contraction[7,41,44]	0.52-0.57
Kyphosis[38]	0.37
Barrel chest[41]	0.62
Thoracic ratio ≥0.9[41]	0.32
Displaced trachea[14]	0.01
Palpation	
Tracheal descent during inspiration[44]	0.62
Laryngeal height ≤5.5 cm[41]	0.59
Impalpable apex beat[14,38]	0.33-0.44
Decreased tactile fremitus[14,43]	0.24-0.86
Increased tactile fremitus[14]	0.01
Subxiphoid point of maximal cardiac impulse[45]	0.30

TABLE 5.1 Interobserver Agreement and Physical Signs—cont'd

Finding (ref)	κ-Statistic[*]
Paradoxical costal margin movement[44,46]	0.56-0.82
Percussion	
Hyperresonant percussion note[14,40,45]	0.26-0.50
Dull percussion note[14,40,43,47]	0.16-0.84
Diaphragm excursion more or less than 2 cm, by percussion[45]	–0.04
Diminished cardiac dullness[45]	0.49
Auscultatory percussion abnormal[43,48]	0.18-0.76
Auscultation	
Reduced breath sound intensity[14,40,41,43,45,47,49,50]	0.16-0.89
Bronchial breathing[14,40]	0.19-0.32
Whispering pectoriloquy[14]	0.11
Reduced vocal resonance[43]	0.78
Crackles[14,47,49,51-54]	0.21-0.65
Wheezes[14,45,47,49,50]	0.43-0.93
Rhonchi[40,50]	0.38-0.55
Pleural rub[14,43]	−0.02-0.51
Special Tests	
Snider test $<$10 cm[45]	0.39
Forced expiratory time[41,45,55,56]	0.27-0.70
Hoover sign[50]	0.74
Wells simplified rule for pulmonary embolism[57]	0.54-0.62
HEART	
Neck Veins	
Neck veins, elevated or normal[51-53,58]	0.08-0.71
Abdominojugular test[58]	0.92
Palpation	
Palpable apical impulse present[59-61]	0.68-0.82
Palpable apical impulse measureable[62]	0.56
Palpable apical impulse displaced lateral to midclavicular line[51,59,60,63]	0.43-0.86
Apical beat normal, sustained, double, or absent[63]	0.88
Percussion	
Cardiac dullness $>$10.5 cm from midsternal line[64,65]	0.57
Auscultation	
S2 diminished or absent, vs. normal[66]	0.54
Third heart sound[51-53,58,67-69]	–0.17-0.84
Fourth heart sound[68,70]	0.15-0.71
Systolic murmur, present or absent[66]	0.19
Systolic murmur radiates to right carotid[66]	0.33
Systolic murmur, long systolic or early systolic[71]	0.78
Murmur intensity (Levine grade)[72]	0.43-0.60
Systolic murmur grade $>$2/6[73]	0.59

Continued

TABLE 5.1 Interobserver Agreement and Physical Signs—cont'd

Finding (ref)	κ-Statistic*
Carotid Pulsation	
Delayed carotid upstroke[66]	0.26
Reduced carotid volume[66]	0.24
ABDOMEN	
Inspection	
Abdominal distension[74,75]	0.35-0.42
Abdominal wall collateral veins, present vs. absent[15]	0.47
Palpation and Percussion	
Ascites[15,17,53]	0.47-0.75
Abdominal tenderness[74-76]	0.31-0.68
Surgical abdomen[75]	0.27
Abdominal wall tenderness test[77,78]	0.52-0.81
Rebound tenderness[74]	0.25
Guarding[74,75]	0.36-0.49
Rigidity[74]	0.14
Abdominal mass palpated[75]	0.82
Palpable spleen[15,17]	0.33-0.75
Palpable liver edge[79]	0.44-0.53
Liver consistency, normal or abnormal[15]	0.4
Liver firm to palpation[80]	0.72
Liver, nodular or not[15]	0.29
Liver, tender or not[17]	0.49
Liver, span >9 cm by percussion[51]	0.11
Spleen palpable or not[81]	0.56-0.70
Spleen percussion sign (Traube), positive or not[82]	0.19-0.41
Abdominal aortic aneurysm, present vs. absent[83]	0.53
Auscultation	
Normal bowel sounds[75]	0.36
EXTREMITIES	
Peripheral Vascular Disease	
Peripheral pulse, present vs. absent[84,85]	0.52-0.92
Peripheral pulse, normal or diminished[84]	0.01-0.15
Cool extremities[53]	0.46
Severity of skin mottling over leg[86,87]	0.87
Diabetic Foot	
Monofilament sensation, normal or abnormal[88-90]	0.48-0.83
Probe-to-bone test[91-93]	0.59-0.84
Edema and Deep Venous Thrombosis	
Dependent edema[51-53]	0.39-0.73
Well pre-test probability for DVT[94,95]	0.74-0.75
Musculoskeletal System—Shoulder	
Shoulder tenderness[96]	0.32
Painful arc[96-99]	0.45-0.64

TABLE 5.1 Interobserver Agreement and Physical Signs—cont'd

Finding (ref)	κ-Statistic*
External rotation of shoulder <45 degrees[96]	0.68
Supraspinatus test (empty can)[96,99,100]	0.44-0.94
Infraspinatus test (resisted external rotation)[96,97]	0.49-0.67
Impingement sign (Hawkins-Kennedy)[96,97,99,100]	0.29-1.0
Drop arm test[96,99]	0.28-0.35
Musculoskeletal System—Hip	
Patrick' test[101]	0.47
Passive internal rotation ≤25 degrees[101]	0.51
Musculoskeletal System—Knee	
Ottawa knee rules[102,103]	0.51-0.77
Knee effusion visible[102,104,105]	0.28-0.59
Knee flexion <90 degrees[102]	0.74
Patellar tenderness[102,104]	0.69-0.76
Head of fibula tenderness[102]	0.64
Inability to bear weight immediately and emergency room after knee injury[102,104]	0.75-0.81
Bony swelling of knee[106]	0.55
Joint line tenderness[105-108]	0.11-0.43
Patellofemoral crepitus[106]	0.24
Mediolateral instability of knee[106]	0.23
McMurray sign[105,108,109]	0.16-0.35
Musculoskeletal System—Ankle	
Inability to walk 4 steps immediately and in emergency room after ankle injury[110,111]	0.71-0.97
Medial malleolar tenderness[111]	0.82
Lateral malleolar tenderness[111]	0.80
Navicular tenderness[111]	0.91
Base of 5th metatarsal tenderness[111]	0.94
Ottawa ankle rule[112]	0.41
Ottawa midfoot rule[112]	0.77
NEUROLOGIC EXAMINATION	
Visual Fields	
Visual fields by confrontation[113]	0.63-0.81
Cranial Nerves	
Pharyngeal sensation, present or absent[114]	1.0
Facial palsy, present or absent[115,116]	0.57
Dysarthria, present or absent[117,118]	0.41-0.77
Water swallow test (50 mL)[119]	0.60
Oxygen desaturation test (for aspiration risk)[119]	0.60
Abnormal tongue strength[117]	0.55-0.63
Motor Examination	
Muscle strength, Medical Research Council (MRC) scale[120-123]	0.69-0.93
Foot tapping test[124]	0.73
Muscle atrophy[125,126]	0.32-0.82

Continued

TABLE 5.1 Interobserver Agreement and Physical Signs—cont'd

Finding (ref)	κ-Statistic*
Spasticity, 6 point scale[127]	0.21-0.61
Rigidity, 4 point scale[128]	0.64
Asterixis[15]	0.42
Tremor[126]	0.74
Pronator drift[129]	0.39
Forearm rolling test[129]	0.73
Sensory Examination	
Light touch sensation, normal, diminished, or increased[125,126]	0.22-0.63
Pain sensation, normal, diminished, or increased[121,125,126]	0.41-0.57
Vibratory sensation, normal or diminished[125,126]	0.28-0.54
Romberg test[126]	0.64
Reflex Examination	
Reflex amplitude, National Institute of Neurological Disorders and Stroke (NINDS) scale[130]	0.51-0.61
Ankle jerk, present or absent[121,131,132]	0.34-0.94
Asymmetric knee jerk[121]	0.42
Babinski response[115,116,124,126,133,134]	0.17-0.60
Finger flexion reflex[135]	0.65
Primitive reflexes, amplitude and persistence[136]	0.46-1.0
Coordination	
Finger-nose test[115,116,126,129]	0.14-0.65
Heel-shin test[126]	0.58
Peripheral Nerve	
Spurling test[137]	0.60
Katz hand diagram[138]	0.86
Flick sign[139]	0.90
Hypalgesia index finger[139]	0.50
Tinel sign[139]	0.47
Phalen sign[139]	0.79
Straight-leg raising test[121,140-144]	0.21-0.80
Crossed-leg raising test[121]	0.49

*Interpretation of the κ-statistic: 0 to 0.2 slight agreement, 0.2 to 0.4 fair agreement, 0.4 to 0.6 moderate agreement, 0.6 to 0.8 substantial agreement, 0.8 to 1.0 almost perfect agreement.

Table 5.2 shows that, for most of our **diagnostic standards**—chest radiography, computed tomography, screening mammography, angiography, magnetic resonance imaging, ultrasonography, endoscopy, and pathology—interobserver agreement is also less than perfect, with κ-statistics similar to those observed with physical signs. Even with laboratory tests, which present the clinician with a single, indisputable number, interobserver disagreement is still possible and even common, simply because the clinician has to interpret the laboratory test's **significance**. For example, in one study of three endocrinologists reviewing the same thyroid function tests and other clinical data of 55 consecutive outpatients with suspected thyroid disease, the endocrinologists disagreed about the final diagnosis 40% of the time.[32] The computerized interpretation of test results performs no better: in a study of pairs

TABLE 5.2 Interobserver Agreement: Diagnostic Standards

Finding (ref)	κ-Statistic*
CHEST RADIOGRAPHY	
Cardiomegaly[58]	0.48
Pulmonary infiltrate[145]	0.38
Pneumonia[146]	0.45
Interstitial edema[58]	0.83
Pulmonary vascular redistribution[58]	0.50
Grading pulmonary fibrosis, 4 point scale[147]	0.45
CONTRAST VENOGRAPHY	
Deep vein thrombosis in leg[148]	0.53
SCREENING MAMMOGRAPHY	
Suspicious lesion, present vs. absent[149]	0.47
DIGITAL SUBTRACTION ANGIOGRAPHY	
Renal artery stenosis[150]	0.65
CORONARY ARTERIOGRAPHY	
Classification of coronary artery lesions[151]	0.33
ARTHROSCOPY	
Inflamed or torn supraspinatus tendon[152]	0.47
COMPUTED TOMOGRAPHY OF HEAD	
Normal or abnormal, patient with stroke[153]	0.60
Lesion on right or left side, patient with stroke[153]	0.65
Mass effect, present or absent[153]	0.52
COMPUTED TOMOGRAPHY OF THE CHEST	
Lung cancer staging[154]	0.40-0.60
Submassive pulmonary embolism present (angiography)[155]	0.47
Coronary lesion on CT coronary angiography[156]	0.57
MAGNETIC RESONANCE IMAGING OF HEAD	
Compatible with multiple sclerosis[157]	0.57-0.87
Pituitary microadenoma present[158]	0.30
MAGNETIC RESONANCE IMAGING OF LUMBAR SPINE	
Intervertebral disc extrusion, protrusion, bulge, or normal[159,160]	0.59
Lumbar nerve root compression[160,161]	0.63-0.83
ULTRASONOGRAPHY	
Calf deep vein thrombosis, present or absent[162]	0.69
Thyroid nodule, present or absent[163,164]	0.57-0.66
Thyroid nodule, cystic or solid[165]	0.64
Goiter is present[34]	0.63
ELECTROCARDIOGRAPHY	
Diagnosis of narrow-complex tachycardia[166]	0.70
ECHOCARDIOGRAPHY	
Severity of valvular regurgitation[167,168]	0.32-0.55
ENDOSCOPY	
Grade of reflux esophagitis[169]	0.55

Continued

TABLE 5.2 Interobserver Agreement: Diagnostic Standards—cont'd

Finding (ref)	κ-Statistic*
PATHOLOGIC EXAMINATION OF LIVER BIOPSY	
Cholestasis[170]	0.40
Alcoholic liver disease[170]	0.49
Cirrhosis[170]	0.59

*Interpretation of the κ-statistic: 0 to 0.2 slight agreement, 0.2 to 0.4 fair agreement, 0.4 to 0.6 moderate agreement, 0.6 to 0.8 substantial agreement, 0.8 to 1.0 almost perfect agreement.

of electrocardiograms taken only 1 minute apart from 92 patients, the computer interpretation was significantly different 40% of the time, even though the tracings showed no change.[171]

By defining abnormal findings precisely, by studying and mastering examination technique, and by observing every detail at the bedside attentively and without bias or distraction, we can minimize interobserver disagreement and make physical diagnosis more precise. It is simply impossible, however, to abstract every detail of clinicians' observations of patients into exact physical signs; in this way, physical diagnosis is no different from any of the other tools we use to categorize disease. So long as both the material and the observers of clinical medicine are human beings, a certain amount of subjectivity will always be with us.

APPENDIX: CALCULATION OF THE κ-STATISTIC

The observations of two observers who are examining the same N patients independently are customarily displayed in a 2 × 2 table, similar to that in Fig. 5.1. Observer A finds the sign to be present in w_1 patients and absent in w_2 patients; observer B finds the sign to be present in y_1 patients and absent in y_2 patients. The two observers agree the sign is present in a patients and absent in d patients. Therefore, the observed agreement (P_O) is

$$P_O = (a + d) / N$$

Calculation of the κ-statistic first requires calculation of the agreement that would have occurred by chance alone. Among all the patients, observer A found the fraction w_1/N to have the sign; therefore, by chance alone, among the y_1 patients with the sign according to observer B, observer A would find the sign in (w_1/N) times y_1 or (w_1y_1/N) patients (i.e., this is the *number* of patients in which both observers agree the sign is present, by chance alone). Similarly, both observers would agree the sign is absent by chance alone in (w_2y_2/N) patients. Therefore, the expected chance agreement (P_E) is their sum, divided by N:

$$P_E = (w_1y_1 + w_2y_2) / N^2$$

This equation shows that agreement by chance alone (P_E) approaches 100% as both w_1 and y_1 approach 0 or N (i.e., both clinicians agree that a finding is rare or that it is very common).

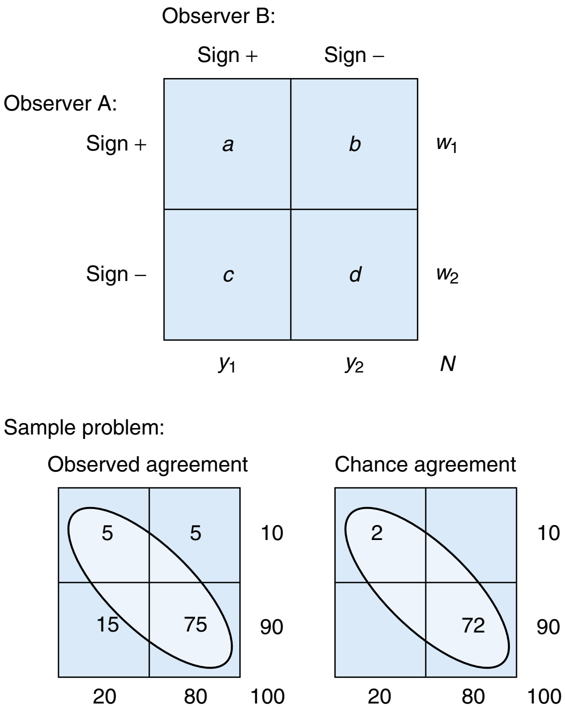

FIG. 5.1 INTEROBSERVER AGREEMENT AND THE κ-STATISTIC. *Top half:* Conventional 2 × 2 table displaying data for calculation of κ-statistic. *Bottom half:* A sample case, in which observed agreement is 80%, chance agreement is 74%, and the κ-statistic is 0.23 (see Appendix for discussion).

The κ-statistic is the increment in observed agreement beyond that expected by chance ($P_O - P_E$), divided by the maximal increment that could have been observed had the observed agreement been perfect ($1 - P_E$):

$$k = \frac{(P_O - P_E)}{(1 - P_E)}$$

For example, Fig. 5.1 depicts the observations of two observers in a study of 100 patients with dyspnea. Both agree the third heart sound is present in 5 patients and absent in 75 patients; therefore simple agreement is (5 + 75)/100 or 0.80. By chance alone, they would have agreed about the sound being present in (10 × 20)/100 patients (i.e., 2 patients) and absent in (90 × 80)/100 patients (i.e., 72 patients); therefore, chance agreement is (2 + 72)/100 patients or 0.74. The κ-statistic for this finding becomes (0.80 – 0.74)/(1 – 0.74) = (0.06)/(0.26) = 0.23.

The references for this chapter can be found on www.expertconsult.com.

This page intentionally left blank

REFERENCES

1. O'Connor DW, Pollitt PA, Hyde JB, et al. The reliability and validity of the mini-mental state in a British community survey. *J Psychiatr Res*. 1989;23(1):87–96.
2. Ainslie NK, Murden RA. Effect of education on the clock-drawing dementia screen in non-demented elderly persons. *J Am Geriatr Soc*. 1993;41:249–252.
3. Inouye SK, Van Dyck CH, Alessi CA, Balkin S, Siegal AP, Horwitz RI. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med*. 1990;113:941–948.
4. Gonzalez M, de Pablo J, Fuente E, et al. Instrument for detection of delirium in general hospitals: adaptation of the confusion assessment method. *Psychosomatics*. 2004;45(5):426–431.
5. Monette J, du Fort GG, Fung SH, et al. Evaluation of the confusion assessment method (CAM) as a screening tool for delirium in the emergency room. *Gen Hosp Psych*. 2001;23:20–25.
6. Fabbri RMA, Moreira MA, Garrido R, Almeida OP. Validity and reliability of the Portuguese version of the confusion assessment method (CAM) for the detection of delirium in the elderly. *Arq Neuropsiquiatr*. 2001;59:175–179.
7. Jones AE, Aborn LS, Kline JA. Severity of emergency department hypotension predicts adverse hospital outcome. *Shock*. 2004;22(5):410–414.
8. Eastlack ME, Arvidson J, Snyder-Mackler L, Danoff JV, McGarvey CL. Interrater reliability of videotaped observational gait-analysis assessments. *Phys Ther*. 1991;71(6):465–472.
9. Verghese J, Lipton RB, Hall CB, Kuslansky G, Katz MJ, Buschke H. Abnormality of gait as a predictor of non-Alzheimer's dementia. *N Engl J Med*. 2002;347:1761–1768.
10. Gjorup T, Hendriksen C, Bugge PM, Jensen AM. Global assessment of patients—a bedside study. II. Interobserver variation and frequency of clinical findings. *J Intern Med*. 1990;228:147–150.
11. Gjorup T, Bugge PM, Hendriksen C, Jensen AM. A critical evaluation of the clinical diagnosis of anemia. *Am J Epidemiol*. 1986;124(4):657–665.
12. Nardone DA, Roth KM, Mazur DJ, McAfee JH. Usefulness of physical examination in detecting the presence or absence of anemia. *Arch Intern Med*. 1990;150:201–204.
13. Sheth TN, Choudhry NK, Bowes M, Detsky AS. The relation of conjunctival pallor to the presence of anemia. *J Gen Intern Med*. 1997;12:102–106.
14. Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet*. 1988;2:873–875.
15. Espinoza P, Ducot B, Pelletier G, et al. Interobserver agreement in the physical diagnosis of alcoholic liver disease. *Dig Dis Sci*. 1987;32(3):244–247.
16. Niederau C, Lange S, Fruhauf M, Thiel A. Cutaneous signs of liver disease: value for prognosis of severe fibrosis and cirrhosis. *Liver Int*. 2008;28(5):659–666.
17. Theodossi A, Knill-Jones RP, Skene A, et al. Inter-observer variation of symptoms and signs in jaundice. *Liver*. 1981;1:21–32.
18. Eaton D, Bannister P, Mulley GP, Connolly MJ. Axillary sweating in clinical assessment of dehydration in ill elderly patients. *Br Med J*. 1994;308:1271.
19. van Genderen ME, Paauwe J, de Jonge J, et al. Clinical assessment of peripheral perfusion to predict postoperative complications after major abdominal surgery early: a prospective observation study in adults. *Crit Care*. 2014;18:R114.
20. Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. *Ann Emerg Med*. 2002;39(3):233–237.
21. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *J Am Med Assoc*. 1993;270:2478–2486.
22. Hla KM, Samsa GP, Stoneking HT, Feussner JR. Observer variability of Osler's maneuver in detection of pseudohypertension. *J Clin Epidemiol*. 1991;44(6):513–518.
23. Belmin J, Visintin JM, Salvatore R, Sebban C, Moulias R. Osler's maneuver: absence of usefulness for the detection of pseudohypertension in an elderly population. *Am J Med*. 1995;98:42–49.
24. Tsapatasaris NP, Napolitana BT, Rothchild J. Osler's maneuver in an outpatient clinic setting. *Arch Intern Med*. 1991;151:2209–2211.

25. Phuong CXT, Nhan NT, Wills B, et al. Evaluation of the World Health Organization standard tourniquet test and a modified tourniquet test in the diagnosis of dengue infection in Viet Nam. *Trop Med Intern Health*. 2002;7(2):125–132.
26. Mayxay M, Phetsouvanh R, Moore CE, et al. Predictive diagnostic value of the tourniquet test for the diagnosis of dengue infection in adults. *Trop Med Int Health*. 2011;16:127–133.
27. Ichhpujani P, Rome JE, Jindal A, et al. Comparative study of 3 techniques to detect a relative afferent pupillary defect. *J Glaucoma*. 2011;20:535–539.
28. Milton RC, Ganley JP, Lynk RH. Variability in grading diabetic retinopathy from stereo fundus photographs: comparison of physician and lay readers. *Br J Ophthalmol*. 1977;61(3):192–201.
29. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs: an extension of the modified Airlie House Classification. ETDRS report number 10. *Ophthalmology*. 1991;98:786–806.
30. Eekhof JAH, de Bock GH, de Laat JAPM, Dap R, Schaapveld K, Springer MP. The whispered voice: the best test for screening for hearing impairment in general practice? *Br J Gen Pract*. 1996;46:473–474.
31. Torres-Russotto D, Landau EM, Harding GW, Bohne BA, Sun K, Sinatra PM. Calibrated finger rub auditory screening test (CALFRAST). *Neurology*. 2009;72:1595–1600.
32. Jarlov AE, Nygaard B, Hegedus L, Hartling SG, Hansen JM. Observer variation in the clinical and laboratory evaluation of patients with thyroid dysfunction and goiter. *Thyroid*. 1998;8(5):393–398.
33. Siminoski K. The rational clinical examination: does this patient have a goiter? *J Am Med Assoc*. 1995;273(10):813–817.
34. Peterson S, Sanga A, Eklof H, et al. Classification of thyroid size by palpation and ultrasonography in field surveys. *Lancet*. 2000;355:106–110.
35. Lindsay KW, Teasdale GM, Knill-Jones RP. Observer variability in assessing the clinical features of subarachnoid hemorrhage. *J Neurosurg*. 1983;58:57–62.
36. Perry JJ, Stiell IG, Sivilotti MLA, et al. Clinical decision rules to rule out subarachnoid hemorrhage for acute headache. *J Am Med Assoc*. 2013;310:1248–1255.
37. Perry JJ, Stiell IG, Sivilotti MLA, et al. High risk clinical characteristics for subarachnoid haemorrhage in patients with acute headache; prospective cohort study. *B Med J*. 2010;341:c5204.
38. Schilling RSF, Hughes JPW, Dingwall-Fordyce I. Disagreement between observers in an epidemiological study of respiratory disease. *Br Med J*. 1955;1:65–68.
39. Pallares-Sanmartin A, Leiro-Fernandez V, Cebreiro TL, Botana-Rial M, Fernandez-Villar A. Validity and reliability of the Schamroth sign for the diagnosis of clubbing. *J Am Med Assoc*. 2010;304(2):159–161.
40. Gjorup T, Bugge PM, Jensen AM. Interobserver variation in assessment of respiratory signs: physicians' guesses as to interobserver variation. *Acta Med Scand*. 1984;216:61–66.
41. de Mattos WLLD, Signori LGH, Borges FK, Bergamin JA, Machado V. Accuracy of clinical examination findings in the diagnosis of COPD. *J Bras Pneumol*. 2009;35(5):404–408.
42. English M, Murphy S, Mwangi I, Crawley J, Peshu N, Marsh K. Interobserver variation in respiratory signs of severe malaria. *Arch Dis Child*. 1995;72:334–336.
43. Kalantri S, Joshi R, Lokhande T, et al. Accuracy and reliability of physical signs in the diagnosis of pleural effusion. *Respir Med*. 2007;101:431–438.
44. Stubbing DG, Mathur PN, Roberts RS, Campbell EJM. Some physical signs in patients with chronic airflow obstruction. *Am Rev Respir Dis*. 1982;125:549–552.
45. Badgett RG, Tanaka DJ, Hunt DK, et al. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med*. 1993;94:188–196.
46. Bruyneel M, Jacob V, Sanida C, Ameye L, Sergysels R, Ninane V. Hoover's sign is a predictor of airflow obstruction severity and is not related to hyperinflation in chronic obstructive pulmonary disease. *Eur J Intern Med*. 2011;22:e115–e118.
47. Mulrow CD, Dolmatch BL, Delong ER, et al. Observer variation in the pulmonary examination. *J Gen Intern Med*. 1986;1:364–367.
48. Nelson RS, Rickman LS, Mathews WC, Beeson SC, Fullerton SC. Rapid clinical diagnosis of pulmonary abnormalities in HIV-seropositive patients by auscultatory percussion. *Chest*. 1994;105:402–407.

49. Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med*. 1993;8:63–68.
50. Garcia-Pachon E. Paradoxical movement of the lateral rib margin (Hoover sign) for detecting obstructive airway disease. *Chest*. 2002;122:651–655.
51. Gadsboll N, Hoilund-Carlsen PF, Nielsen GG, et al. Symptoms and signs of heart failure in patients with myocardial infarction: reproducibility and relationship to chest X-ray, radionuclide ventriculography and right heart catheterization. *Eur Heart J*. 1989;10:1017–1028.
52. Maestre A, Gil V, Gallego J, Aznar J, Mora A, Martin-Hidalgo A. Diagnostic accuracy of clinical criteria for identifying systolic and diastolic heart failure: cross-sectional study. *J Eval Clin Pract*. 2009;15:55–61.
53. Chaudhry A, Singer AJ, Chohan J, Russo V, Lee C. Interrater reliability of hemodynamic profiling of patients with heart failure in the ED. *Am J Emerg Med*. 2008;26:196–201.
54. Quinn JV, Stiell I, McDermott DA, Sellers KL, Kohn MA, Wells GA. Derivation of the San Francisco syncope rule to predict patients with short-term serious outcomes. *Ann Emerg Med*. 2004;43:224–232.
55. Schapira RM, Schapira MM, Funahashi A, McAuliffe TL, Varkey B. The value of the forced expiratory time in the physical diagnosis of obstructive airways disease. *J Am Med Assoc*. 1993;270(6):731–736.
56. Badgett R, Tanaka D. The diagnostic value of the forced expiratory time. *J Am Med Assoc*. 1994;271(1):25.
57. Wolf SJ, McCubbin TR, Feldhaus KM, Faragher JP, Adcock DM. Prospective validation of Wells criteria in the evaluation of patients with suspected pulmonary embolism. *Ann Emerg Med*. 2004;44:503–510.
58. Butman SM, Ewy GA, Standen JR, Kern KB, Hahn E. Bedside cardiovascular examination in patients with severe chronic heart failure: importance of rest or inducible jugular venous distension. *J Am Coll Cardiol*. 1993;22(4):968–974.
59. O'Neill TW, Smith M, Barry M, Graham IM. Diagnostic value of the apex beat. *Lancet*. 1989;1(8635):410–411.
60. O'Neill TW, Barry MA, Smith M, Graham IM. Diagnostic value of the apex beat. *Lancet*. 1989;2(8661):499.
61. Mulkerrin E, Saran R, Dewar R, Harding JR, Bayer AJ, Finucane P. The apex cardiac beat: not a reliable clinical sign in elderly patients. *Age Ageing*. 1991;20(4):304–306.
62. Dans AL, Bossone EF, Guyatt GH, Fallen EL. Evaluation of the reproducibility and accuracy of apex beat measurement in the detection of echocardiographic left ventricular dilation. *Can J Cardiol*. 1995;11(6):493–497.
63. Ehara S, Okuyama T, Shirai N, et al. Comprehensive evaluation of the apex beat using 64-slice computed tomography: impact of left ventricular mass and distance to chest wall. *J Cardiol*. 2010;55:256–265.
64. Heckerling PS, Wiener SL, Moses VK, Claudio J, Kushner MS, Hand R. Accuracy of precordial percussion in detecting cardiomegaly. *Am J Med*. 1991;91:328–334.
65. Heckerling PS, Wiener SL, Wolfkiel CJ, et al. Accuracy and reproducibility of precordial percussion and palpation for detecting increased left ventricular end-diastolic volume and mass: a comparison of physical findings and ultrafast computed tomography of the heart. *J Am Med Assoc*. 1993;270(16):1943–1948.
66. Etchells E, Glenns V, Shadowitz S, Bell C, Siu S. A bedside clinical prediction rule for detecting moderate or severe aortic stenosis. *J Gen Intern Med*. 1998;13:699–704.
67. Ishmail AA, Wing S, Ferguson J, Hutchinson TA, Magder S, Flegel KM. Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure. *Chest*. 1987;91(6):870–873.
68. Lok CE, Morgan CD, Ranganathan N. The accuracy and interobserver agreement in detecting the "gallop sounds" by cardiac auscultation. *Chest*. 1998;114:1283–1288.
69. Tribouilloy CM, Enriquez-Sarano M, Mohty D, et al. Pathophysiologic determinants of third heart sounds: a prospective clinical and Doppler echocardiography study. *Am J Med*. 2001;111:96–102.
70. Meyers DG, Porter IT, Schneider KA, Maksoud AR. Correlation of an audible fourth heart sound with level of diastolic dysfunction. *Am J Med Sci*. 2009;337(3):165–168.

71. Forssell G, Jonasson R, Orinius E. Identifying severe aortic valvular stenosis by bedside examination. *Acta Med Scand*. 1985;218:397–400.
72. Keren R, Tereschuk M, Luan X. Evaluation of a novel method for grading heart murmur intensity. *Arch Pediatr Adolesc Med*. 2005;159:329–334.
73. Reichlin S, Dieterle T, Camli C, Leimenstroll B, Schoenenberger RA, Martina B. Initial clinical evaluation of cardiac systolic murmurs in the ED by noncardiologists. *Am J Emerg Med*. 2004;22:71–75.
74. Bjerregaard B, Brynitz S, Holst-Christensen J, et al. The reliability of medical history and physical examination in patients with acute abdominal pain. *Methods Inform Med*. 1983;22:15–18.
75. Pines J, Pines LU, Hall A, Hunter J, Srinivasan R, Ghaemmaghami C. The interrater variation of ED abdominal examination findings in patients with acute abdominal pain. *Am J Emerg Med*. 2005;23:483–487.
76. Priebe WM, DaCosta LR, Beck IT. Is epigastric tenderness a sign of peptic ulcer disease? *Gastroenterol*. 1982;82:16–19.
77. Srinivasan R, Greenbaum DS. Chronic abdominal wall pain: a frequently overlooked problem. Practical approach to diagnosis and management. *Am J Gastroenterol*. 2002;97(4):824–830.
78. Takada T, Ikusaka M, Ohira Y, Noda K, Tsukamoto T. Diagnostic usefulness of Carnett's test in psychogenic abdominal pain. *Intern Med (Tokyo)*. 2011;50:213–217.
79. Joshi R, Singh A, Jajoo N, Pai M, Kalantri SP. Accuracy and reliability of palpation and percussion for detecting hepatomegaly: a rural hospital-based study. *Indian J Gastroenterol*. 2004;23:171–173.
80. Tine F, Caltagirone M, Camma C, et al. Clinical indicants of compensated cirrhosis: a prospective study. In: Dianzani MU, Gentilini P, eds. *Chronic Liver Damage: proceedings of the Annual Meeting of the Italian National Programme on Liver Cirrhosis. San Miniato, Italy 11-13 January 1990*. Amsterdam: Excerpta Medica; 1990:187–198.
81. Barkun AN, Camus M, Green L, et al. The bedside assessment of splenic enlargement. *Am J Med*. 1991;91:512–518.
82. Barkun AN, Camus M, Meagher T, et al. Splenic enlargement and Traube's space: how useful is percussion? *Am J Med*. 1989;87:562–566.
83. Fink HA, Lederle FA, Roth CS, Bowles CA, Nelson DB, Haas MA. The accuracy of physical examination to detect abdominal aortic aneurysm. *Arch Intern Med*. 2000;160:833–836.
84. Myers KA, Scott DF, Devine TJ, Johnston AH, Denton MJ, Gilfillan IS. Palpation of the femoral and popliteal pulses: a study of the accuracy as assessed by agreement between multiple observers. *Eur J Vasc Surg*. 1987;1:245–249.
85. Brearley S, Shearman CP, Simms MH. Peripheral pulse palpation: an unreliable physical sign. *Ann R Coll Surg Engl*. 1992;74:169–171.
86. Ait-Oufella H, Lemoinne S, Boelle PY, et al. Mottling score predicts survival in septic shock. *Intensive Care Med*. 2011;37:801–807.
87. Coudroy R, Jamet A, Frat JP, et al. Incidence and impact of skin mottling over the knee and its duration on outcome in critically ill patients. *Intensive Care Med*. 2015;41:452–459.
88. Diamond JE, Mueller MJ, Delitto A, Sinacore DR. Reliability of diabetic foot evaluation. *Phys Ther*. 1989;69(10):797–802.
89. Smieja M, Hunt DL, Edelman D, Etchells E, Cornuz J, Simel DL. Clinical examination for the detection of protective sensation in the feet of diabetic patients. *J Gen Intern Med*. 1999;14:418–424.
90. Edelman D, Sanders LJ, Pogach L. Reproducibility and accuracy among primary care providers of a screening examination for foot ulcer risk among diabetic patients. *Prevent Med*. 1998;27:274–278.
91. Lozano RM, Montesinos JVB, Fernández MLG, Jiménez SG, Hernández DM, Jurado MAG. Validating the probe-to-bone test and other tests for diagnosing chronic osteomyelitis in the diabetic foot. *Diabetes Care*. 2010;33:2140–2145.
92. Álvaro-Afonso FJ, Lázaro-Martínez JL, Aragón-Sánchez J, García-Morales E, García-Álvarez Y, Molines-Barroso RJ. Inter-observer reproducibility of diagnosis of diabetic

foot osteomyelitis based on a combination of probe-to-bone test and simple radiography. *Diabetes Res Clin Pract*. 2014;105:e3–e5.

93. García Morales E, Lázaro-Martínez JL, Aragón-Sánchez FJ, Cecilia-Matilla A, Beneit-Montesinos JV, González Jurado MA. Inter-observer reproducibility of probing to bone in the diagnosis of diabetic foot osteomyelitis. *Diabet Med*. 2011;28:1238–1240.
94. Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet*. 1997;350:1795–1798.
95. Dewar C, Corretge M. Interrater reliability of the Wells score as part of the assessment of DVT in the emergency department: agreement between consultant and nurse practitioner. *Emerg Med J*. 2008;25:407–410.
96. Ostor AJK, Richards CA, Prevost AT, Hazleman BL, Speed CA. Interrater reproducibility of clinical tests for rotator cuff lesions. *Ann Rheum Dis*. 2004;63:1288–1292.
97. Michener LA, Walsworth MK, Doukas WC, Murphy KP. Reliability and diagnostic accuracy of 5 physical examination tests and combination of tests for subacromial impingement. *Arch Phys Med Rehabil*. 2009;90:1898–1903.
98. Nomden JG, Slagers AJ, Bergman GJD, Winters JC, Kropmans TJB, Dijkstra PU. Interobserver reliability of physical examination of shoulder girdle. *Man Ther*. 2009;14:152–159.
99. Nanda R, Gupta S, Kanapathipillai P, Liow RYL, Rangan A. An assessment of the inter examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity. *Eur J Orthop Surg Traumatol*. 2008;18:495–500.
100. Johansson K, Ivarson S. Intra- and interexaminer reliability of four manual shoulder maneuvers used to identify subacromial pain. *Man Ther*. 2009;14:231–239.
101. Sutlive TG, Lopez HP, Schnitker DE, et al. Development of a clinical prediction rule for diagnosing hip osteoarthritis in individuals with unilateral hip pain. *J Orthop Sports Phys Ther*. 2008;38(9):542–550.
102. Stiell IG, Greenberg GH, Wells GA, et al. Prospective validation of a decision rule for the use of radiography in acute knee injuries. *J Am Med Assoc*. 1996;275:611–615.
103. Cheung TC, Tank Y, Breederveld RS, Tuinebreijer WE, Lange-de Klerk ESM, Derksen RJ. Diagnostic accuracy and reproducibility of the Ottawa knee rule vs. the Pittsburgh decision rule. *Am J Emerg Med*. 2013;31:641–645.
104. Stiell IG, Greenberg GH, Wells GA, et al. Derivation of a decision rule for the use of radiography in acute knee injuries. *Ann Emerg Med*. 1995;26(4):405–413.
105. Dervin GF, Stiell IG, Wells GA, Rody K, Grabowski J. Physicians' accuracy and interrator reliability for the diagnosis of unstable meniscal tears in patients having osteoarthritis of the knee. *Can J Surg*. 2001;44(4):267–274.
106. Cushnagham J, Cooper C, Dieppe P, Kirwan J, McAlindon T, McCrae F. Clinical assessment of osteoarthritis of the knee. *Ann Rheum Dis*. 1990;49:768–770.
107. Snoeker BAM, Lindeboom R, Zwinderman AH, Vincken PWJ, Jansen JA, Lucas C. Detecting meniscal tears in primary care: reproducibility and accuracy of 2 weight-bearing tests and 1 non-weight-bearing test. *J Ortho Sports Phys Ther*. 2015;45:693–702.
108. Galli M, Ciriello V, Menghi A, Aulisa AG, Rabini A, Marzetti E. Joint line tenderness and McMurray tests for the detection of meniscal lesions: what is their real diagnostic value? *Arch Phys Med Rehabil*. 2013;94:1126–1131.
109. Evans PJ, Bell GD, Frank C. Prospective evaluation of the McMurray test. *Am J Sports Med*. 1993;21(4):604–608.
110. Stiell IG, Greenberg GH, McKnight RD, et al. Decision rules for the use of radiography in acute ankle injuries: refinement and prospective validation. *J Am Med Assoc*. 1993;269:1127–1132.
111. Springer BA, Arciero RA, Tenuta JJ, Taylor DC. A prospective study of modified Ottawa ankle rules in a military population: interobserver agreement between physical therapists and orthopaedic surgeons. *Am J Sports Med*. 2000;28(6):864–868.
112. Derkson RJ, Bakker FC, Geervliet PC, et al. Diagnostic accuracy and reproducibility in the interpretation of Ottawa ankle and foot rules by specialized emergency nurses. *Am J Emerg Med*. 2005;23:725–729.
113. Kerr NM, Chew SSL, Eady EK, Gamble GD, Danesh-Meyer HV. Diagnostic accuracy of confrontation visual field tests. *Neurology*. 2010;74:1184–1190.

114. Davies AE, Kidd K, Stone SP, MacMahon J. Pharyngeal sensation and gag reflex in healthy subjects. *Lancet*. 1995;345:487–488.
115. Hansen M, Christensen PB, Sindrup SH, Olsen NK, Kristensen O, Friis ML. Interobserver variation in the evaluation of neurological signs: patient-related factors. *J Neurol*. 1994;241:492–496.
116. Hansen M, Sindrup SH, Christensen PB, Olsen NK, Kristensen O, Friis ML. Interobserver variation in the evaluation of neurologic signs: observer dependent factors. *Acta Neurol Scand*. 1994;90:145–149.
117. McCullough GH, Wertz RT, Resenbek JC, Mills RH, Ross KB, Ashford JR. Inter- and intrajudge reliability of a clinical examination of swallowing in adults. *Dysphagia*. 2000;15:58–67.
118. Hand PJ, Haisma JA, Kwan J, et al. Interobserver agreement for the bedside clinical assessment of suspected stroke. *Stroke*. 2006;37:776–780.
119. Lim SHB, Lieu PK, Phua SY, et al. Accuracy of bedside clinical methods compared with fiberoptic endoscopic examination of swallowing (FEES) in determining the risk of aspiration in acute stroke patients. *Dysphagia*. 2001;16:1–6.
120. Segatore M. Determining the interrater reliability of motor power assessments using a spinal cord testing record. *J Neurosci Nurs*. 1991;23(4):220–223.
121. Vroomen PCAJ, de Krom MC, Knottnerus JA. Consistency of history taking and physical examination in patients with suspected lumbar nerve root involvement. *Spine*. 2000;25(1):91–97.
122. Brandsma JW, Schreuders TAR, Birke JA, Piefer A, Oostendorp R. Manual muscle strength testing: intraobserver and interobserver reliabilities for the intrinsic muscles of the hand. *J Hand Ther*. 1995;8:185–190.
123. Jeon CH, Chung NS, Lee YS, Son KH, Kim JH. Assessment of hip abductor power in patients with foot drop. *Spine*. 2013;38(3):257–263.
124. Miller TM, Johnston SC. Should the Babinski sign be part of the routine neurologic examination? *Neurology*. 2005;65:1165–1168.
125. Viikari-Juntura E. Interexaminer reliability of observations in physical examinations of the neck. *Phys Ther*. 1987;67(10):1526–1532.
126. Thaller M, Hughes T. Inter-rater agreement of observable and elicitable neurological signs. *Clin Med*. 2014;14:264–267.
127. Haas BM, Bergstrom E, Jamous A, Bennie A. The inter rater reliability of the original and of the modified Ashworth scale for the assessment of spasticity in patients with spinal cord injury. *Spin Cord*. 1996;34:560–564.
128. van Dillen L, Roach KE. Interrater reliability of a clinical scale of rigidity. *Phys Ther*. 1988;68(11):1679–1681.
129. Amer M, Hubert G, Sullivan SJ, Herbison P, Franz EA, Hammond-Tooke GD. Reliability and diagnostic characteristics of clinical tests of upper limb motor function. *J Clin Neurosci*. 2012;19:1246–1251.
130. Litvan I, Mangone CA, Werden W, et al. Reliability of the NINDS myotatic reflex scale. *Neurology*. 1996;47:969–972.
131. O'Keeffe STO, Smith T, Valacio R, Jack CIA, Playfer JR, Lye M. A comparison of two techniques for ankle jerk assessment in elderly subjects. *Lancet*. 1994;344:1619–1620.
132. Clarke CE, Davies P, Wilson T, Nutbeam T. Comparison of the tendon and plantar strike methods of eliciting the ankle reflex. *J Neurol Neurosurg Psychiatry*. 2005;74:1351–1352.
133. Maher J, Reilly M, Daly L, Hutchinson M. Plantar power: reproducibility of the plantar response. *Br Med J*. 1992;304:482.
134. Singerman J, Lee L. Consistency of the Babinski reflex and its variants. *Eur J Neurol*. 2008;15(9):960–964.
135. Annaswamy TM, Sakai T, Goetz LL, Pacheco FM, Ozarkar T. Reliability and repeatability of the Hoffmann sign. *PM R*. 2012;4:498–503.
136. Vreeling FW, Jolles J, Verhey FRJ, Houx PJ. Primitive reflexes in healthy, adult volunteers and neurological patients: methodological issues. *J Neurol*. 1993;240:495–504.
137. Wainner RS, Fritz JM, Irrgang JJ, Boninger ML, Delitto A, Allison S. Reliability and diagnostic accuracy of the clinical examination and patient self-report measures for cervical radiculopathy. *Spine*. 2003;28(1):52–63.

138. Calfee RP, Dale AM, Ryan D, Descatha A, Franzblau A, Evanoff B. Performance of simplified scoring systems for hand diagrams in carpal tunnel syndrome screening. *J Hand Surg Am*. 2012;37:10–17.
139. Wainner RS, Fritz JM, Irrgang JJ, Delitto A, Allison S, Boninger ML. Development of a clinical prediction rule for the diagnosis of carpal tunnel syndrome. *Arch Phys Med Rehabil*. 2005;86:609–618.
140. McCombe PF, Fairbank JCT, Cockersole BC, Pynsent PB. Reproducibility of physical signs in low-back pain. *Spine*. 1989;14(9):909–918.
141. Van den Hoogen HJM, Koes BW, Deville W, Van Eijk JTM, Bouter LM. The interobserver reproducibility of Lasegue's sign in patients with low back pain in general practice. *Br J Gen Pract*. 1996;46:727–730.
142. Poiraudeau S, Foltz V, Drape JL, et al. Value of the bell test and the hyperextension test for diagnosis in sciatica associated with disc herniation: comparison with Lasegue's sign and the crossed Lasegue's sign. *Rheumatology*. 2001;40:460–466.
143. Rabin A, Gerszte PC, Karausky P, Bunker CH, Potter DM, Welch WC. The sensitivity of the seated straight-leg raise test compared with the supine straight-leg test in patients presenting with magnetic resonance imaging evidence of lumbar nerve root compression. *Arch Phys Med Rehabil*. 2007;88:840–843.
144. Walsh J, Hall T. Agreement and correlation between the straight leg raise and slump tests in subjects with leg pain. *J Manip Physiol Ther*. 2009;32:184–192.
145. Albaum MN, Hill LC, Murphy M, et al. Interobserver reliability of the chest radiography in community-acquired pneumonia. *Chest*. 1996;110:343–350.
146. van Vugt SF, Verhij TJM, de Jong PA, et al. Diagnosing pneumonia in patients with acute cough: clinical judgment compared to chest radiography. *Eur Respir J*. 2013;42:1076–1082.
147. Baughman RP, Shipley RT, Loudon RG, Lower EE. Crackles in interstitial lung disease: comparison of sarcoidosis and fibrosing alveolitis. *Chest*. 1991;100:96–101.
148. Illescas FF, Lerclerc J, Resenthall L, et al. Interobserver variability in the interpretation of contrast venography, technetium-99m red blood cell venography and impedance plethysmography for deep vein thrombosis. *J Can Assoc Radiol*. 1990;41:264–269.
149. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(22):1493–1499.
150. DeVries AR, Engels PHC, Overtoom TT, Saltzherr TP, Geyskes BG. Interobserver variability in assessing renal artery stenosis by digital subtraction angiography. *Diagn Images Clin Med*. 1984;53:277–281.
151. Herrman JP, Azar A, Umans VA, Boersma E, von Es GA, Serruys PW. Inter- and intra-observer variability in the qualitative categorization of coronary angiograms. *Int J Cardiovasc Imaging*. 1996;12:21–30.
152. Mohtadi NGH, Jager FL, Sasyniuk TM, Hollinshead RM, Fick GH. The reliability between surgeons comparing arthroscopic and video evaluation of patients with shoulder impingement syndrome. *Arthroscopy*. 2004;20(10):1055–1062.
153. Shinar D, Gross CR, Hier DB, et al. Interobserver reliability in the interpretation of computed tomographic scans of stroke patients. *Arch Neurol*. 1987;44:149–155.
154. Webb WR, Sarin M, Zerhouni EA, Heelan RT, Glazer GM, Gatsonis C. Interobserver variability in CT and MR staging of lung cancer. *J Comput Assist Tomogr*. 1993;17(6):841–846.
155. Costantino G, Norsa AH, Amadori R, et al. Interobserver agreement in the interpretation of computed tomography in acute pulmonary embolism. *Am J Emerg Med*. 2009;27:1109–1111.
156. Øvrehus KA, Marwa M, Bøtker HE, Achenbach S, Nørgaard BL. Reproducibility of coronary opaque detection and characterization using low radiation dose coronary computed tomographic angiography in patients with intermediate likelihood of coronary artery disease (ReSCAN study). *Int J Cardiovasc Imaging*. 2012;28:889–899.
157. Barkhof F, Filippi M, van Waesberghe JH, Campi A, Miller DH, Ader HJ. Interobserver agreement for diagnostic MRI criteria in suspected multiple sclerosis. *Neuroradiol*. 1999;41:347–350.

158. Bahurel-Barrera H, Assie G, Silvera S, Bertagna X, Costa J, Legmann P. Inter- and intra-observer variability in detection and progression assessment with MRI of microadenoma in Cushing's disease patients followed up after bilateral adrenalectomy. *Pituitary*. 2008;11:263–269.
159. Jensen MC, Brant-Zawadzki MN, Obuchowski N, Modic MT, Malkasian D, Ross JS. Magnetic resonance imaging of the lumbar spine in people without back pain. *N Engl J Med*. 1994;331(2):69–73.
160. Kuijper B, Beelen A, van der Kallen BF, et al. Interobserver agreement on MRI evaluation of patients with cervical radiculopathy. *Clin Radiol*. 2011;66:25–29.
161. Vroomen PC, de Krom MC, Wilmink JT, Kester AD, Knottnerus JA. Diagnostic value of history and physical examination in patients suspected of lumbosacral nerve root compression. *J Neurol Neurosurg Psychiatry*. 2002;72:630–634.
162. Atri M, Herba MJ, Reinhold C, et al. Accuracy of sonography in the evaluation of calf deep vein thrombosis in both postoperative surveillance and symptomatic patients. *Am J Roentgenol*. 1996;166:1361–1367.
163. Jarlov AE, Nygard B, Hegedus L, Karstrup S, Hansen JM. Observer variation in ultrasound assessment of the thyroid gland. *Br J Radiol*. 1993;66:625–627.
164. Schneider AB, Bekerman C, Leland J, et al. Thyroid nodules in the follow-up of irradiated individuals: comparison of thyroid ultrasound with scanning and palpation. *J Clin Endocrinol Metab*. 1997;82(12):4020–4027.
165. Park CS, Kim SH, Jung SL, et al. Observer variability in the sonographic evaluation of thyroid nodules. *J Clin Ultrasound*. 2010;38:287–293.
166. González-Torrecilla E, Almendral J, Arenal A, et al. Combined evaluation of bedside clinical variables and the electrocardiogram for the differential diagnosis of paroxysmal atrioventricular reciprocating tachycardias in patients without pre-excitation. *J Am Coll Cardiol*. 2009;53:2353–2358.
167. Biner S, Rafique A, Rafii F, et al. Reproducibility of proximal isovelocity surface area, vena contracta, and regurgitant jet area for assessment mitral regurgitation severity. *J Am Coll Cardiol Imging*. 2010;3:325–343.
168. Grant ADM, Thavendiranathan P, Rodriguez LL, Kwon D, Marwick TH. Development of a consensus algorithm to improve interobserver agreement and accuracy in the determination of tricuspid regurgitation severity. *J Am Soc Echocardiogr*. 2014;27:277–284.
169. Bytzer P, Havelund T, Moeller Hansen J. Interobserver variation in the endoscopic diagnosis of reflux esophagitis. *Scand J Gastroenterol*. 1993;28:119–125.
170. Theodossi A, Skene AM, Portmann B, et al. Observer variation in assessment of liver biopsies including analysis by kappa statistics. *Gastroenterology*. 1980;79:232–241.
171. Spodick DH, Bishop RL. Computer treason: intraobserver variability of an electrocardiographic computer system. *Am J Cardiol*. 1997;80(1):102–103.